EXPERIMENTAL AND THEORETICAL ANALYSIS OF HYBRIDIZATION

Simone Linz[1 2], Katherine St John[3 4], and Charles Semple[5]

Department of Mathematics and Statistics
University of Canterbury
Private Bag 4800
Christchurch, New Zealand

Report Number: UCDMS2006/4 MAY 2006

Heinrich-Heine-University, Düsseldorf, Germany, linz@cs.uni-dusseldorf.de
Center for Integrative Bioinformatics, Vienna, Austria
Dept. of Math and Computer Science, Lehman College-City U. of New York, USA, tjohn@lehman.cuny.edu
Dept. of Computer Science, CUNY Graduate Center
Biomathematics Research Centre, Department of Mathematics and Statistics, U. of Canterbury, New Zealand, c.semple@math.canterbury.ac.nz

Experimental & Theoretical Analysis of Hybridization

Simone Linz[1,2], Katherine St. John[3,4], and Charles Semple[5]

[1] Heinrich-Heine-University, Düsseldorf, Germany, linz@cs.uni-duesseldorf.de
[2] Center for Integrative Bioinformatics, Vienna, Austria,
[3] Dept. of Math & Computer Science, Lehman College–City U. of New York, USA, stjohn@lehman.cuny.edu
[4] Dept. of Computer Science, CUNY Graduate Center
[5] Biomathematics Research Centre, Department of Mathematics and Statistics, U. of Canterbury, New Zealand, c.semple@math.canterbury.ac.nz

Abstract. We develop new heuristics and an exact algorithm for calculating the amount of hybridization between two rooted binary phylogenetic trees. Calculating the minimum number of hybridization events is NP-hard, but essential to understanding the modeling of reticulation processes such as hybridization, horizontal gene transfer, and recombination. We give new lower bounds for the hybridization number that are very useful in limiting search times for exact answers and in conjunction with existing upper bounds to "sandwich" the true answer. We analyze the algorithms experimentally on both biological and simulated data.

1 Introduction

The analysis and understanding of reticulation in the evolutionary history of a collection of present-day species is now a prominent and central area of study in phylogenetics [4, 15, 20]. Instead of the usual tree-like processes of evolution, reticulation processes such as hybridization, horizontal gene transfer, and recombination result in evolution behaving in a non-tree-like way as some species are a mixture of genes derived from different ancestors. Thus, for certain groups of species whose evolutionary past includes reticulation (e.g. particular groups of plants and fish), a rooted acyclic digraph is a better representation of their evolutionary history.

Despite the occurrence of reticulation events, it is commonly accepted that such events are relatively rare and so a fundamental problem for biologists studying the ancestral history of a group of species whose past includes reticulation is the following: given a collection of rooted binary phylogenetic trees on a set of species that correctly represent the tree-like evolution of different parts of their genomes, what is the smallest number of reticulation events needed to explain the evolutionary history of the species under consideration. The mathematical formalization of this problem results in an NP-hard problem even when the initial collection consists of two rooted binary phylogenetic trees [10]. However, for this two-tree problem, there is a recent theoretical result of Baroni [3] that is the basis of a divide-and-conquer type approach for finding the exact solution. In the context of this particular problem, the smallest number is often referred to as the "hybridization number" and, for two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$, it is denoted by $h(\mathcal{T}, \mathcal{T}')$ (see Section 2).

Historically, the "rooted subtree prune and regraft" (rSPR) distance between $\mathcal{T}$ and $\mathcal{T}'$ has often been used as a replacement for $h(\mathcal{T}, \mathcal{T}')$. Denoted by $d_{\text{rSPR}}(\mathcal{T}, \mathcal{T}')$, this distance, roughly speaking, equates to the minimum number of subtrees that must be "moved" to transform $\mathcal{T}$ into $\mathcal{T}'$ (see Section 2). Like computing $h(\mathcal{T}, \mathcal{T}')$, computing $d_{\text{rSPR}}(\mathcal{T}, \mathcal{T}')$ is also NP-hard [9]. The reason for using the rSPR distance as a replacement for the hybridization number is that $h(\mathcal{T}, \mathcal{T}') = 1$ if and only if $d_{\text{rSPR}}(\mathcal{T}, \mathcal{T}') = 1$, and if one interpolates this for $h(\mathcal{T}, \mathcal{T}') = k$, then, intuitively, it appears that $d_{\text{rSPR}}(\mathcal{T}, \mathcal{T}') = k$. However, one can find explicit examples that show that the difference between these two values can be arbitrarily large [6]. Nevertheless, in practice, it appears that the rSPR distance provides a reasonable lower bound to the hybridization number. Furthermore, there is a fixed-parameter algorithm for computing $d_{\text{rSPR}}(\mathcal{T}, \mathcal{T}')$ [9].

In this paper, we analyze the divide-and-conquer approach for finding the exact hybridization number of two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ and develop new polynomial-time heuristics for lower bounds to the hybridization number. The lower bound heuristics complement past work on upper bounds for the hybridization number and can be used to reduce the search time for exact algorithms. Our experiments of the implemented algorithms, along with heuristics of ours and others design, are on both simulated and biological data sets.

Much of the related past work focuses on complexity and approximation results for rSPR distances [2, 8, 9, 14, 23]. In addition to the theoretical work described above, there has been related past work on finding upper bounds for the hybridization number of phylogenetic trees, and we have included much of this in our experimental analysis. This includes the RIATA-HGT algorithm of Nakhleh *et al.* [19] which greedily constructs the upper bound by finding maximum agreement subtrees (described in more detail in Section 3). Beiko and Hamilton [7] very recently developed Efficient Evaluation of Edit Paths (EEEP) which bounds the number of rSPR moves (the "edit path") between two trees, subject to biologically reasonable constraints (see Section 3). Hallet *et al.* [13, 1] developed LATTRANS for special cases of the problem. Unfortunately, like [19], we were not able to run the code and could not include it in our experimental analysis.

2 Background

In this section, we formally define the rSPR distance and hybridization number of two rooted binary phylogenetic trees as well as the concept of an agreement forest which is central to many of the results described in this paper. Terminology and notation follow [27]. Recall that, for $|X| \geq 2$, a rooted binary phylogenetic X-tree $\mathcal{T}$ is a rooted tree whose root has degree two and all other interior vertices have degree three, and whose leaf set is X. If $|X| = 1$, then, for completeness, the rooted tree consisting of an isolated vertex labeled by the element in X is defined to be a rooted binary phylogenetic tree. In the definitions that follow for "rSPR distance" and "agreement forests", we regard the root of $\mathcal{T}$ as a vertex ρ at the end of a pendant edge (called the *root edge*) adjoined to the original root. To illustrate, see Fig. 1.

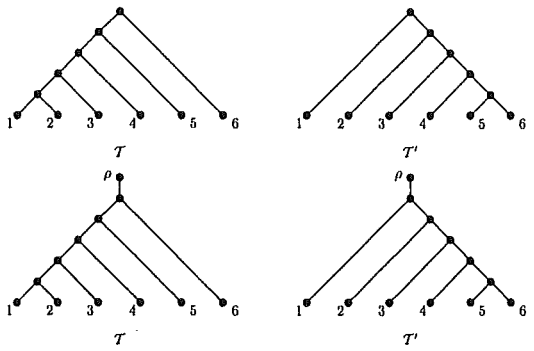


Fig. 1. Two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ without (above) and with (below) their roots labelled.

2.1 rSPR Distance

Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree and let $e = \{u, v\}$ be an edge of $\mathcal{T}$ that is not the root edge, where u is the vertex that is in the path from the root of $\mathcal{T}$ to v. Let $\mathcal{T}'$ be the rooted binary phylogenetic tree obtained from $\mathcal{T}$ by deleting e and then adjoining a new edge f between v and the component C_u that contains u as follows. Create a new vertex u' which subdivides an edge in C_u, and adjoin f between u' and v, and then contract the degree-two vertex u. We say that $\mathcal{T}'$ has been obtained from $\mathcal{T}$ by a **rooted subtree prune and regraft** (rSPR) operation. We define the rSPR **distance** between two arbitrary rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ to be the minimum number of rooted subtree prune and regraft operations that is required to transform $\mathcal{T}$ into $\mathcal{T}'$. Denoted by $d_{\text{rSPR}}(\mathcal{T}, \mathcal{T}')$, it is well-known that this distance is well-defined. As an example, in Fig. 2, each of $\mathcal{T}'$ and $\mathcal{T}''$ are obtained from $\mathcal{T}$ by one rSPR operation.

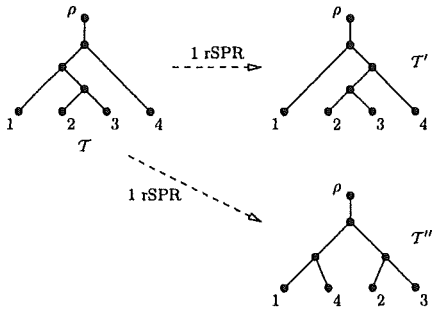


Fig. 2. Both $\mathcal{T}'$ and $\mathcal{T}''$ are obtained from $\mathcal{T}$ by a single rSPR operation.

It is shown in [9] that the optimization problem MINIMUM RSPR of computing the rSPR distance is NP-hard. However, in the same paper, it also shown to be FPT. Both results rely on the equivalence between the rSPR distance and a version of agreement forests described in Section 2.3.

2.2 Hybridization Number

For a digraph D and a vertex v of D, we denote the in-degree and out-degree of v by $d^-(v)$ and $d^+(v)$, respectively. A **hybridization network** (on X) is a rooted acyclic digraph with root ρ in which

(i) X is the set of vertices of out-degree zero,
(ii) $d^+(\rho) \geq 2$, and
(iii) for all vertices v with $d^+(v) = 1$, we have $d^-(v) \geq 2$.

For completeness, if $|X| = 1$, then the digraph consisting of an isolated vertex labelled by the element in X is also defined to be a hybridization network on X. Biologically speaking, the set X is a collection of present-day species. Vertices of in-degree at least two (called *hybridization vertices*) represent reticulation events, which we generically refer to as hybridization events, and correspond to an exchange of genetic information between hypothetical ancestors. The **hybridization number** of $\mathcal{H}$, denoted $h(\mathcal{H})$, is

$$h(\mathcal{H}) = \sum_{v \neq \rho} (d^-(v) - 1),$$

where ρ denotes the root of $\mathcal{H}$. Noting that every vertex apart from the root has at least one parent, the value "$d^-(v)-1$" represents the number of additional parents of v. A hybridization network $\mathcal{H}$ is shown in Fig. 3 with $h(\mathcal{H}) = 2$.

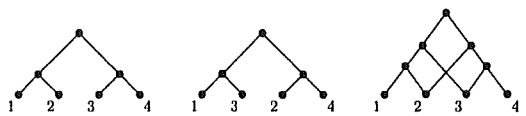


Fig. 3. Two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$, and a hybridization network $\mathcal{H}$ that displays them.

Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree and let $\mathcal{H}$ be a hybrid phylogeny on X. We say that $\mathcal{H}$ **displays** $\mathcal{T}$ if $\mathcal{T}$ can be obtained from a rooted *subtree* of $\mathcal{H}$ by contracting degree-two vertices. For example, in Fig. 3, $\mathcal{H}$ displays both $\mathcal{T}$ and $\mathcal{T}'$. For two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$, we set

$$h(\mathcal{T}, \mathcal{T}') = \min\{h(\mathcal{H}) : \mathcal{H} \text{ displays } \mathcal{T} \text{ and } \mathcal{T}'\}.$$

The optimization problem MINIMUM HYBRIDIZATION of computing $h(\mathcal{T}, \mathcal{T}')$ is NP-hard [10]. As for MINIMUM rSPR, this hardness is based on a particular type of agreement forest which we describe in §2.3. We end this section by noting that, loosely speaking, the rSPR distance can differ from the hybridization number for a pair of trees since a sequence of rSPR operations that transforms one tree into the other may result in unwanted directed cycles in the canonical hybridization network that is constructed from these operations (see [26] for further details).

2.3 Agreement Forests

An essential tool in the analysis of the rSPR distance and hybridization number is the use of agreement forests. Originally developed by [14], variants correspond to both the rSPR distance and hybridization number. We briefly describe the relevant details needed for the algorithms in this paper.

Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree and let X' be a subset of X. The minimal rooted subtree of $\mathcal{T}$ that connects the vertices of $\mathcal{T}$ labelled by the elements of X' is denoted by $\mathcal{T}(X')$. Furthermore, the **restriction** of $\mathcal{T}$ to X', denoted by $\mathcal{T}|X'$, is the rooted binary phylogenetic tree that is obtained from $\mathcal{T}(X')$ by contracting any non-root vertices of degree two.

Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Recall that, for the definition of an agreement forest, we regard the root of both $\mathcal{T}$ and $\mathcal{T}'$ as a vertex ρ at the end of a pendant edge adjoined to the original root. Furthermore, for the definitions in this section, we also regard ρ as part of the label sets of $\mathcal{T}$ and $\mathcal{T}'$, thus we view both label sets as $X \cup \{\rho\}$. An **agreement forest** for $\mathcal{T}$ and $\mathcal{T}'$ is a collection $\{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$, where $\mathcal{T}_\rho$ is a rooted tree whose label set $\mathcal{L}_\rho$ includes ρ and $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k$ are rooted binary phylogenetic trees with label sets $\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_k$, respectively, such that the following properties are satisfied:

(i) The label sets $\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_k$ partition $X \cup \{\rho\}$.
(ii) For all $i \in \{\rho, 1, 2, \ldots, k\}$, $\mathcal{T}_i \cong \mathcal{T}|\mathcal{L}_i \cong \mathcal{T}'|\mathcal{L}_i$.
(iii) The trees in $\{\mathcal{T}(\mathcal{L}_i) : i \in \{\rho, 1, 2, \ldots, k\}\}$ and $\{\mathcal{T}'(\mathcal{L}_i) : i \in \{\rho, 1, 2, \ldots, k\}\}$ are vertex disjoint rooted subtrees of $\mathcal{T}$ and $\mathcal{T}'$, respectively.

It is easily seen that if $\mathcal{F}$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, then, up to contracting non-root vertices of degree two, $\mathcal{F}$ can be obtained from each of $\mathcal{T}$ and $\mathcal{T}'$ by deleting $|\mathcal{F}| - 1$ edges. An agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ is a **maximum-agreement forest** if,

amongst all agreement forests for $\mathcal{T}$ and $\mathcal{T}'$, it has the smallest number of components, in which case we denote the value of k by $m(\mathcal{T}, \mathcal{T}')$. To illustrate, Fig. 4 shows two agreement forests $\mathcal{F}_1$ and $\mathcal{F}_2$ for $\mathcal{T}$ and $\mathcal{T}'$ in Fig. 1, where the root ρ of each of $\mathcal{T}$ and $\mathcal{T}'$ is adjoined to the original root as described above.

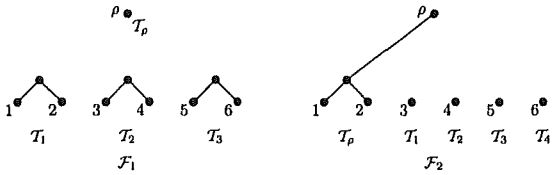


Fig. 4. A maximum-agreement forest $\mathcal{F}_1$ for the two trees $\mathcal{T}$ and $\mathcal{T}'$ in Fig. 1, and a maximum-acyclic-agreement forest $\mathcal{F}_2$ for $\mathcal{T}$ and $\mathcal{T}'$.

The equivalence between the rSPR distance and the size of a maximum-agreement forest is given in the following Theorem:

Theorem 1. [9] *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Then $d_{\text{rSPR}}(\mathcal{T}, \mathcal{T}') = m(\mathcal{T}, \mathcal{T}')$*

To illustrate Theorem 1, it is easily checked that $\mathcal{F}_1$ in Fig. 4 is a maximum-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ in Fig. 1, and so, by Theorem 1, $d_{\text{rSPR}}(\mathcal{T}, \mathcal{T}') = 3$.

A similar equality to that in Theorem 1 can also be obtained for $h(\mathcal{T}, \mathcal{T}')$ by placing a restriction on the agreement forest, and thus avoiding the unwanted directed cycles referred to at the end of the last subsection. Let $\mathcal{F} = \{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$ be an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Let $G_\mathcal{F}$ be the directed graph whose vertex set is $\mathcal{F}$ and for which $(\mathcal{T}_i, \mathcal{T}_j)$ is an arc precisely if $i \neq j$ and either

(I) the root of $\mathcal{T}(\mathcal{L}_i)$ is an ancestor of the root of $\mathcal{T}(\mathcal{L}_j)$, or
(II) the root of $\mathcal{T}'(\mathcal{L}_i)$ is an ancestor of the root of $\mathcal{T}'(\mathcal{L}_j)$.

Since $\mathcal{F}$ is an agreement forest, the roots of $\mathcal{T}(\mathcal{L}_i)$ and $\mathcal{T}(\mathcal{L}_j)$, and the roots of $\mathcal{T}'(\mathcal{L}_i)$ and $\mathcal{T}'(\mathcal{L}_j)$ are not the same. We say that $\mathcal{F}$ is an **acyclic-agreement forest** if $G_\mathcal{F}$ is acyclic. (Note that, in [6], the authors use "good" instead of "acyclic" as we have used here.) Furthermore, if $\mathcal{F}$ contains the smallest number of components over all acyclic-agreement forests for $\mathcal{T}$ and $\mathcal{T}'$, we say that $\mathcal{F}$ is a **maximum-acyclic-agreement forest** for $\mathcal{T}$ and $\mathcal{T}'$, in which case we denote this value of k by $m_a(\mathcal{T}, \mathcal{T}')$. Observe that $m_a(\mathcal{T}, \mathcal{T}') = 0$ if and only if, up to isomorphism, $\mathcal{T}$ and $\mathcal{T}'$ are identical. To illustrate, Fig. 5 contains the graphs $G_{\mathcal{F}_1}$ and $G_{\mathcal{F}_2}$, where $\mathcal{F}_1$ and $\mathcal{F}_2$ are the agreement forests shown in Fig. 4. Since $G_{\mathcal{F}_1}$ contains a directed cycle, $\mathcal{F}_1$ is not an acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ in Fig. 1, however, as $G_{\mathcal{F}_2}$ is acyclic, $\mathcal{F}_2$ is a such a forest.

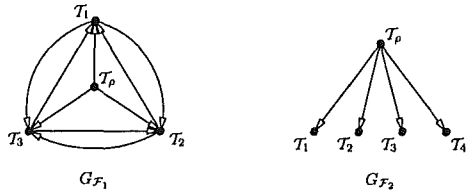


Fig. 5. The graphs $G_{\mathcal{F}_1}$ and $G_{\mathcal{F}_2}$, where $\mathcal{F}_1$ and $\mathcal{F}_2$ are as shown in Fig. 4.

The equivalence between the hybridization number and maximum-acyclic-agreement forests is given in the next theorem.

Theorem 2. [6] *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Then $h(\mathcal{T}, \mathcal{T}') = m_a(\mathcal{T}, \mathcal{T}')$.*

Referring back to the last example, it is easily checked that $\mathcal{F}_2$ is a maximum-acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, and so $h(\mathcal{T}, \mathcal{T}') = 4$.

3 Methods

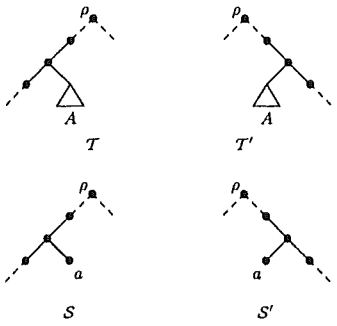
We use both heuristics and exact algorithms to calculate the hybridization number. Our exact algorithms rely on recent theoretical developments for calculating the rSPR distance and the hybridization number of two rooted binary phylogenetic trees. We have developed heuristics that give lower bounds for the hybridization number and use those developed by others to give upper bounds.

3.1 Exact Algorithms

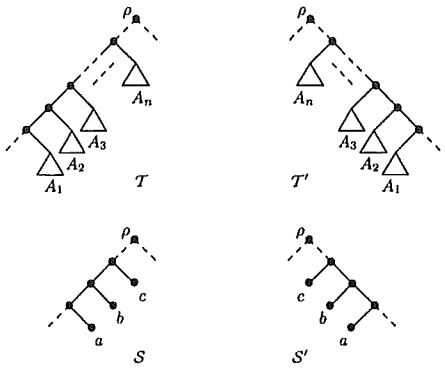
rSPR Distance: For computing the rSPR distance between two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$, we employ the FPT result of [9]. To show FPT, the authors use the following two rules to "kernalize" the trees. These rules are modified versions of the two reduction rules in [2]. Each of the figures illustrate the corresponding rule, where $\mathcal{S}$ and $\mathcal{S}'$ are the

trees resulting from a single application of the rule, and the "new" labels are a, and a, b, c, respectively.

Rule 1 Replace any pendant subtree that occurs identically in both trees by a single leaf with a new label:



Rule 2 Replace any chain of pendant subtrees that occurs identically and with the same orientation relative to the root in both trees by three new leaves with new labels correctly orientated to preserve the direction of the chain:



The next proposition is shown in [9].

Proposition 1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees, and let $\mathcal{S}$ and $\mathcal{S}'$ be the rooted binary phylogenetic trees that are obtained from $\mathcal{T}$ and $\mathcal{T}'$ by applying either Rule 1 or Rule 2. Then*

$$d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') = d_{\mathrm{rSPR}}(\mathcal{S},\mathcal{S}').$$

As a consequence of Proposition 1, one can repeatedly apply Rules 1 and 2 to $\mathcal{T}$ and $\mathcal{T}'$ to obtain two rooted binary phylogenetic X'-trees $\mathcal{S}$ and $\mathcal{S}'$ with the property that $d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') = d_{\mathrm{rSPR}}(\mathcal{S},\mathcal{S}')$. If this is done so that no further reductions are possible, then $|X'| \leq 28 d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}')$ [9]. This inequality is the basis for the theoretical justification that computing rSPR distance is FPT, where the parameter is d_{rSPR}. This leads to the obvious fixed-parameter algorithm for computing the rSPR distance between two trees, which we call RSPRDIST. It takes the reduced trees $\mathcal{S}$ and $\mathcal{S}'$ and exhaustively searches the possible agreement forests. Due to the tractability and equivalence between the rSPR distance and the size of a maximum agreement forest, this is an exact algorithm for computing the rSPR distance between two trees. Moreover, it also provides a lower bound for the hybridization number of two trees.

Hybridization Number: An efficient way to divide-and-conquer the calculation of the hybridization number for two rooted binary phylogenetic trees is to use the following result in [3] (also see [5]). For a rooted phylogenetic X-tree $\mathcal{T}$, a subset of X is called a **cluster** if it is the set of descendant leaves of some vertex in $\mathcal{T}$.

Proposition 2. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees, and suppose that A is a cluster of both $\mathcal{T}$ and $\mathcal{T}'$. Let $\mathcal{T}_a$ and $\mathcal{T}'_a$ be the rooted binary phylogenetic trees obtained from $\mathcal{T}$ and $\mathcal{T}'$, respectively, by replacing $\mathcal{T}|A$ and $\mathcal{T}'|A$ with a single vertex a, where $a \notin X$. Then*

$$h(\mathcal{T},\mathcal{T}') = h(\mathcal{T}|A,\mathcal{T}'|A) + h(\mathcal{T}_a,\mathcal{T}'_a).$$

Observe that it is an immediate consequence of Proposition 2 that, as the hybridization number of two, up to isomorphism, identical trees is zero, we can apply the reduction implied by Rule 1 to $\mathcal{T}$ and $\mathcal{T}'$ to obtain two rooted binary phylogenetic trees $\mathcal{S}$ and $\mathcal{S}'$ such that $h(\mathcal{T},\mathcal{T}') = h(\mathcal{S},\mathcal{S}')$. However, we point out here that the analogous reduction implied by Rule 2 does not work for the hybridization number (see [26]).

Proposition 2 is the basis for the following divide-and-conquer algorithm for computing the hybridization number of two rooted binary phylogenetic trees.

The non-polynomial part of the algorithm is finding an appropriate agreement-forest in Step 4. A first-up approach would be to exhaustively delete edges from both $\mathcal{T}$ and $\mathcal{T}'$, and then see if the resulting forests are the same. However, a much faster approach is to (exhaustively) delete edges from just

Algorithm: HYBRIDNUMBER($\{\mathcal{T}, \mathcal{T}'\}$)
Input: Two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$
Output: The value of $h(\mathcal{T}, \mathcal{T}')$.

1. Set $\mathcal{T}_0 = \mathcal{T}$ and $\mathcal{T}'_0 = \mathcal{T}'$, and set $i = 1$
2. Repeatedly apply Rule 1 to $\mathcal{T}_{i-1}$ and $\mathcal{T}'_{i-1}$ until completely reduced and set $\mathcal{S}_{i-1}$ and $\mathcal{S}'_{i-1}$ to be the resulting trees, resp. If $\mathcal{S}_{i-1}$ and $\mathcal{S}'_{i-1}$ both consist of a single vertex, then go to Step 7.
3. Find a minimal common cluster, W_{i-1} of $\mathcal{S}_{i-1}$ and $\mathcal{S}'_{i-1}$ with $|W_{i-1}| > 2$.
4. Find a maximum-acyclic-agreement forest F_{i-1} for $\mathcal{S}_{i-1}|W_{i-1}$ and $\mathcal{S}'_{i-1}|W_{i-1}$.
5. Set $\mathcal{T}_i$ and $\mathcal{T}'_i$ to be the trees obtained from $\mathcal{S}_{i-1}$ and $\mathcal{S}'_{i-1}$, respectively, by replacing $\mathcal{S}_{i-1}|W_{i-1}$ and $\mathcal{S}'_{i-1}|W_{i-1}$ with a single vertex w_{i-1}.
6. Increment i by 1 and return to Step 2.
7. Output the sum $|\mathcal{F}_0| - 1 + |\mathcal{F}_1| - 1 + \cdots + |\mathcal{F}_{i-1}| - 1$.

one of the trees, $\mathcal{T}$ say, to obtain a forest $\mathcal{F} = \{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$ and then see if the collection $\{\mathcal{T}'(\mathcal{L}_\rho), \mathcal{T}'(\mathcal{L}_1), \mathcal{T}'(\mathcal{L}_2), \ldots, \mathcal{T}'(\mathcal{L}_k)\}$ of subtrees of $\mathcal{T}'$ are vertex disjoint. If no, then $\mathcal{F}$ is not an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. On the other hand, if yes, then $\mathcal{F}$ is such a forest. Of course, one also needs to check that an agreement forest is acyclic, which can be done quickly. This approach can also be used in the FPT algorithm described in the previous section.

Note that the analog of Proposition 2 does not hold for computing the rSPR distance between two trees. In particular, using the notation and terminology in the statement of this theorem, it is shown in [9] that

$$\begin{aligned} d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') &\leq d_{\mathrm{rSPR}}(\mathcal{T}|A, \mathcal{T}'|A) + d_{\mathrm{rSPR}}(\mathcal{T}_a, \mathcal{T}'_a) \\ &\leq d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') + 1 \end{aligned}$$

with both inequalities being sharp. For further details see [9, 26].

3.2 Heuristics

Counting Iterations: The algorithm HYBRIDNUMBER provides a fast lower bound for the hybridization number if one ignores the computationally expensive Step 4. In particular, as each iteration contributes at least one to this number, the total number of iterations is a lower bound for it. We refer to the resulting algorithm as HYBRIDAPPROX.

RIATA-HGT: A more sophisticated polynomial-time heuristic has been described by Nakhleh *et al.* [19]. In this heuristic, they find what is effectively an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ by repeatedly finding a maximum-agreement subtree of two trees to decompose $\mathcal{T}$ and $\mathcal{T}'$ appropriately. The resulting agreement forest gives an upper bound to rSPR distance. It is not known if this agreement forest is acyclic, so, it may underestimate the hybridization number (although this does not seem to happen in practice, see Section 5).

EEEP: Beiko and Hamilton [7] have developed a tool, EEEP, for calculating exact and upper bound heuristics for the related problem of determining minimal "edit paths" between two trees. The edit path between two trees is a set of rSPR moves that transforms one tree into the other. They require that the paths be acyclic, yielding edit paths that correspond to the hybridization number. Since reconstructed phylogenetic trees are often unrooted, EEEP that only the species or reference tree is rooted. This can give lower scores than the hybridization number of the corresponding rooted trees.

Counting Cherries: Recalling that the rSPR distance and hybridization number is preserved under Rule 1, a simple and fast heuristic that provides a lower bound for the rSPR distance, and thereby the hybridization number, of two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ involves counting the total number of "cherries" in the two trees obtained from $\mathcal{T}$ and $\mathcal{T}'$ by repeatedly applying Rule 1 until it can no longer be applied. This heuristic, AVERAGECHERRIES is stated as Proposition 3.

A **cherry** of a rooted binary phylogenetic tree $\mathcal{T}$ is a pendant subtree with exactly two leaves. If the two leaves are a and b say, then we denote the cherry by (a, b). Furthermore, the total number of cherries of $\mathcal{T}$ is denoted by $c(\mathcal{T})$.

Proposition 3. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees for which no reduction results in the application of Rule 1. Then*

$$\frac{c(\mathcal{T}) + c(\mathcal{T}')}{2} \leq d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}').$$

Proof. Let $n = |X|$ and let $k = c(\mathcal{T}) + c(\mathcal{T}')$. The proof is by induction on n. First observe that, as $\mathcal{T}$ and $\mathcal{T}'$ are reduced under Rule 1, $\mathcal{T}$ and $\mathcal{T}'$ have no

common cherries. Now $n = 1$ if and only if, up to isomorphism, $\mathcal{T}$ and $\mathcal{T}'$ are identical, in which case $k = 0$ and the result holds. Therefore we may assume that $\mathcal{T}$ and $\mathcal{T}'$ are not identical, in which case $n \geq 3$ and the rSPR distance is non-zero. Indeed, if $k = 3$, then, as each tree has exactly one cherry, the result also holds.

Now assume that $n \geq 4$, and that the result holds whenever the size of X is less than n. Let $\mathcal{F}$ be a maximum-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Without loss of generality, we may assume that $c(\mathcal{T}) \geq c(\mathcal{T}')$. Let $\mathcal{F}$ be a maximum-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Then either (i) every cherry (a, b) in $\mathcal{T}$ has the property that a and b are contained in the same label set of a tree in $\mathcal{F}$ or (ii) there exists one such cherry with a in the label set of one tree in $\mathcal{F}$ and b in the label set of another tree in $\mathcal{F}$. If (i) holds, then, as $\mathcal{T}$ and $\mathcal{T}'$ have no cherries in common, it is easily seen that all such cherries give rise to at least one edge in $\mathcal{T}'$ that must be deleted to obtain $\mathcal{F}$. Moreover, an easy check shows that these edges can be chosen so that they are pairwise distinct. Since $c(\mathcal{T}) \geq c(\mathcal{T}')$, the proposition holds.

Suppose that (ii) holds. Then, without loss of generality, we may assume that the edge connecting a to $\mathcal{T}$ is deleted in creating $\mathcal{F}$. Let $\mathcal{S}$ and $\mathcal{S}'$ be the trees obtained from $\mathcal{T}$ and $\mathcal{T}'$, respectively, by deleting a and then applying Rule 1 until it can no longer be applied. Note that $\mathcal{S}$ and $\mathcal{S}'$ have no cherries in common. Since $\mathcal{F}\backslash a$ is an agreement forest for $\mathcal{T}\backslash a$ and $\mathcal{T}'\backslash a$ and since Rule 1 preserves the rSPR distance, it follows that $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') > d_{\mathrm{rSPR}}(\mathcal{S}, \mathcal{S}')$. Furthermore, as we removed only a, it is easily seen that $c(\mathcal{T}) - c(\mathcal{S}) \in \{0, 1\}$ and $c(\mathcal{T}') - c(\mathcal{S}') \in \{0, 1\}$. In other words,

$$c(\mathcal{S}) + c(\mathcal{S}') \geq c(\mathcal{T}) + c(\mathcal{T}') - 2.$$

Since the size of the leaf sets of $\mathcal{S}$ and $\mathcal{S}'$ are less than n, it now follows by the induction assumption that

$$\begin{aligned} d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') &\geq d_{\mathrm{rSPR}}(\mathcal{S}, \mathcal{S}') + 1 \geq \frac{c(\mathcal{S}) + c(\mathcal{S}')}{2} + 1 \\ &\geq \frac{c(\mathcal{T}) + c(\mathcal{T}') - 2}{2} + 1 = \frac{c(\mathcal{T}) + c(\mathcal{T}')}{2} \end{aligned}$$

□

It is shown in [8] that reductions by Rule 1 and 2 can be done in linear time. Since the algorithm AVERAGECHERRIES visits each edge of the reduced trees at most once, it can also be implemented in linear time.

3.3 Computational Resources

We obtained the RIATA-HGT code from the authors [19] and the EEEP code from the authors [7]. All other code was written by the authors in perl and java. The java code uses the code base of TreeJuxtaposer software package [18] (freely available at `http://olduvai.sourceforge.net`).

The data was analyzed on three separate linux clusters. The helix cluster at the Allan Wilson Centre for Molecular Ecology and Evolution (AWC) in New Zealand is a distributed-memory parallel machine (a Beowulf cluster) with 65 nodes (130 processors), running the Linux operating system and communicating with the MPI protocol. The nodes are dual processors Athlon MP-2100 with 1 GB memory. Wildebeest is a 132-processor Beowulf cluster, located at City University of New York, and administered by the Research Computing Group at the Graduate Center of CUNY. The nodes are Athlon 2000+ dual processors (1.664GHz) with 1GB memory. The cluster at the University of Düsseldorf is administered by the Department of Bioinformatics. It is a 36 node cluster with two different types of CPUs: AMD Opteron 246 (2000 MHz) and AMD Opteron 244 (1800 MHz). Each node has a dual processor with between 2 and 8 GB RAM.

4 Data

4.1 The Grass (*Poaceae*) Dataset

Although the number of hybridization events and its impact on evolution is still discussed controversially [22] the occurrence of such events in plants is generally accepted. In 1996, Ellstrand *et al.* examined the frequency of spontaneous hybridization events in plants [11]. They analyzed the distribution of hybridization in 5 different biosystematic floras and, for 4 of those, the *Poaceae* family is among the 6 families with the highest number of hybrids (between 19 and 45 depending on the flora). Therefore, we can assume that the grass (*Poaceae*) dataset, provided by the Grass Phylogeny Working Group [12], is adequate to study the number of hybridization events.

The mentioned dataset consists of sequence data for six loci, three nuclear and three chloroplast ones. The genes coded in the nuclear DNA are NADH dehydrogenase, subunit F (*ndhF*), granule bound starch synthase I (*waxy*) and the internal transcribed spacer (ITS) whereas the chloroplast ones are ribulose 1,5-bisphosphate carboxylase/oxygenase, large subunit

(*rbcL*), RNA polymerase II, β' subunit (*rpoC*) and phytochrome B (*phyB*). More detailed characteristics about this dataset with an overall number of 66 taxa are summerized in Table 1. The dataset also includes composite taxa which are represented by sequences from several species and genera respectively.
For each gene, a phylogenetic tree was reconstructed, using the fastDNAml program [21]. The resulting trees were provided by Heiko Schmidt who has also analyzed this dataset [25]. To calculate the rSPR distance and hybridization number for each of the 15 pairs of gene trees we had to restrict the trees to the overlapping taxa (Table 2). As there was a polytomy at the root of all trees these were resolved in an arbitrary way, namely (a,b,c) to (a,(b,c)).

Table 1. The *Poaceae* dataset.

loci	sequence origin	#sequences	alignment length
ndhF	chloroplast	65	2210
phyB	nucleus	40	1182
rbcL	chloroplast	37	1344
rpoC	chloroplast	34	777
waxy	nucleus	19	773
ITS	nucleus	47	322

Table 2. Number of overlapping taxa for each pair.

	ndhF	*phyB*	*rbcL*	*rpoC*	*waxy*	ITS
ndhF	-	40	36	34	19	46
phyB		-	21	21	14	30
rbcL			-	26	12	29
rpoC				-	10	31
waxy					-	15
ITS						-

4.2 Simulated Datasets

We generated two different simulated datasets: one for testing overall performance and one for comparing the exact algorithm with the upper bound heuristics.

For each dataset, we calculated the performance of the heuristics: AVERAGECHERRIES, HYBRIDAPPROX, RIATA-HGT, and the exact algorithm HYBRIDNUMBER.

Random Trees: For the first dataset, we simulated closely related trees within a small, medium, and large distances (rSPR distance of less than 3, 5, and 10, resp.) We randomly generated a "species tree" (under the Yule-Harding distribution, using the `r8s` program [24]) and 10 "gene trees" within a fixed rSPR distance. We ran each algorithm on species tree paired with each of its 10 related gene trees, and compared accuracy and running time between the algorithms.

We sampled trees at sizes of $10, 20, \ldots, 100$ leaves. Due to the size of treespace $((2n-3)!!$ rooted binary phylogenetic tree shapes on n leaves), it is not possible to fairly sample the entire input space. In order to obtain statistically robust results, we followed the advice of McGeoch [16] and Moret [17] and used a number of *runs*, each composed of a number of *trials* (a trial is a single comparison between a species tree and related gene tree), computed the average for each run, and studied the mean and standard deviation over the runs of these events. This is done to minimize the effects of the pseudorandom generator that is used to generate the trees. In this approach, each run is an independent pseudorandom stream, and as such, the average of the runs will tend towards the true value, even with the limitations of the pseudorandom generator.

Caterpillars: Our second simulated dataset is a known difficult family of trees and is used for testing the upper bound heuristics. Let $(a_1, a_2, a_3, \ldots, a_n)$ represent the rooted binary phylogenetic tree: $(((\cdots(((a_1, a_2), a_3), \cdots, a_{n-2}), a_{n-1}), a_n)$ (a "caterpillar"). It is easy to see that the caterpillars $(1, 2, 3, \ldots, n)$ and $(4, 5, 6, \ldots, n, 1, 2, 3)$ have hybridization number 3, but calculating this can be difficult. Note that the reduction rules of [9] do not apply to these pairs, since for Rule 1, there are no common cherries; and the hybridization number is not preserved under Rule 2. So, despite the small hybridization number, these pairs can have arbitrarily large size. This dataset consists of the pairs of caterpillars for $n = 6, 7, 8, \ldots, 20$.

5 Results

In general, the methods did well in either time or accuracy, but no method excelled at both. Due to the hardness of the problem, all methods have difficulties at calculating or estimating the hybridization number for non-trivial examples with more than 70 leaves. The exact algorithms also had difficulties for

pairs of trees with smaller leaf sets. For the more difficult examples, the gap between the upper and lower bound heuristics was quite large. This suggests that a combination of the current methods would be the best approach for obtaining the hybridization number (see Section 6).

5.1 Biological Data

The complete output from our experiment on biological data is included in Table 3. An abbreviated version for easier reading is presented as a graph in Fig. 6. The table includes the output along with the running time for the four algorithms included in the analysis. For HYBRIDAPPROX and RIATA-HGT, we also include analysis, when known, on the clusters returned by the algorithms: the size of the cluster and the number of events reported, resp. The distribution of leaves and events in the clusters are included since they indicate the size of the subproblems. When there is a larger number of clusters with evenly distributed size and events, it suggests that a "divide-and-conquer" approach could improve the results.

The graph in Fig. 6 shows the completed results on each of the 15 pairs of trees. There are large differences between the upper and lower bounds tends to grow with tree size.

Given the hardness of the problem instances, many of the exact cases did not finish after 2 weeks of running time (indicated by a 'x' for running time (RT) in Table 3). Since those algorithms exhaustively try all possibilities for each number of cuts, they give lower bounds, even when not run to completion. Those lower bounds are included in the table, preceded by a $\geq$ sign. A positive note is that HybridNumber algorithm did produce results quickly on some of the instances. For example, it gave the exact answer of 8 hybridizations events in 141 seconds on two trees of 30 taxa each. While we have seen that it is not always going to work well in practice, there are some non-trivial instances for which it works exceptionally well.

Due to the late availability of EEEP (we discovered it only a week before the deadline), it did not run to completion on the biological data has not been included in the results.

The results of RIATA-HGT depend on the ordering of gene and species tree. For each case, we report the ordering that gives the minimum number of horizontal gene transfer events.

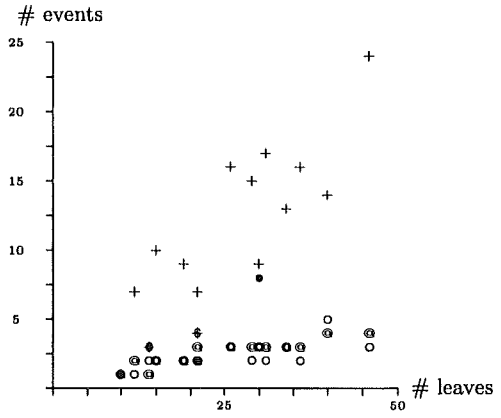


Fig. 6. Grass dataset: Number of hybrid events reported by AVERAGECHERRIES (@), HYBRIDAPPROX (o), HYBRIDNUMBER (•), RIATA-HGT (+), and RSPRDIST ($).

5.2 Simulated Data

To test the effectiveness of the algorithms, we looked at several different sets of simulated data. These included randomly generated datasets which have closely related trees and tailored datasets from the literature to illustrate differences between the methods.

Random Trees: HYBRIDNUMBER, RSPRDIST and RIATA-HGT ran into memory constraints that made calculating some of the instances impossible. For HYBRIDNUMBER and RSPRDIST, we stopped runs if there was no answer after 2 weeks of computation time. We did not have a sufficient number finish to include these in the analysis below. RIATA-HGT always crashed with trees of more than 70 taxa but ran well on smaller trees. The lower bound heuristics always ran to completion and are included in the analysis.

Note that for even very large trees, the HYBRIDAPPROX does poorly bounding the hybridization number, compared to the simple AVERAGECHERRIES heuristic. This contrasts with their performance on the biological data, where HYBRIDAPPROX did better.

Table 3. Results for the Grass Dataset

dataset (#taxa)	RSPRDIST		AVERAGECHERRIES	HYBRIDNUMBER		HYBRIDAPPROX		RIATA-HGT	
	#events	RT [s]		#events	RT [s]	#events	#leaves/cluster	#events	#events/cluster
ndhF phyB (40)	≥ 3	x	4	≥ 4	x	5	3,3,3,4,15	14	1,1,1,1,10
ndhF rbcL (36)	≥ 4	x	3	4	x	2	7,15	16	x
ndhF rpoC (34)	≥ 4	x	3	4	x	3	3,4,16	13	x
ndhF waxy (19)	≥ 5	x	2	5	x	2	4,11	9	x
ndhF ITS (46)	≥ 3	x	4	4	x	3	3,10,20	24	x
phyB rbcL (21)	4	3975	2	4	0	2	1,5,6	4	x
phyB rpoC (21)	≥ 4	x	3	≥ 5	x	2	4,11	7	x
phyB waxy (14)	3	28	1	3	0	2	1,4,4	3	x
phyB ITS (30)	≥ 4	x	3	8	141	3	1,3,7,9	9	5,1,4,0
rbcL rpoC (26)	≥ 4	x	3	≥ 4	x	3	3,4,16	16	x
rbcL waxy (12)	≥ 5	x	2	5	x	1	11	7	x
rbcL ITS (29)	≥ 4	x	3	4	x	2	3,21	15	x
rpoC waxy (10)	1	0	1	1	0	1	1,4	1	x
rpoC ITS (31)	≥ 4	x	3	≥ 4	x	2	4,19	17	x
waxy ITS (15)	≥ 5	x	2	≥ 5	x	2	3,12	10	x

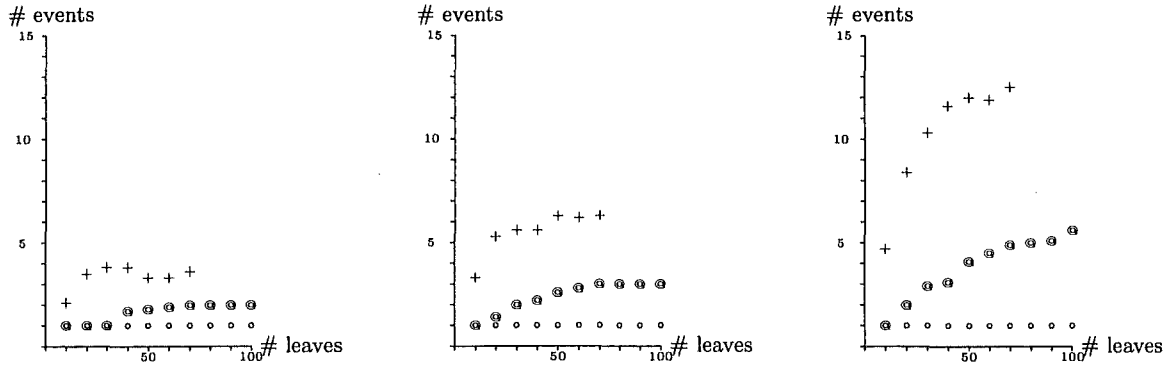


Fig. 7. Number of hybrid events reported by RIATA=HGT (+), HybridApprox (o), and AverageCherries (@) for 3 move, 5 move, and 10 move datasets.

Caterpillars: For each pair of caterpillars (described in Section 4), we ran the exact algorithm HYBRIDNUMBER, as well as the upper bound heuristics, RIATA-HGT and EEEP. By construction, the true answer for every pair is 3. The exact algorithm gave the correct answer, but took increasingly more time as the caterpillars grew in size (see Fig. 8). The heuristics both overestimated the hybridization number, and this overestimate grew linearly with the size of the caterpillars. The RIATA-HGT heuristic had the best running times, staying nearly constant; the other two programs running times grew exponentially as the size of the caterpillars grew. Due to time constraints, we ran EEEP only with the default settings. It includes heuristics (called "ratchets") which should greatly improve its performance.

6 Discussion

All the algorithms tested have strengths, and their combined use is currently the best approach to calculating the hybridization numbers for biological datasets. The lower bound approximations developed in this paper give good starting points for exact searches and work well in conjunction with the upper bound approximations for "sandwiching" the true answer.

HYBRIDAPPROX shows the promise of divide-and-conquer approach and gives hope that the exact answer can be found. It does much better on biological data than the simulated datasets. Surprisingly, given the FPT result for RSPRDIST, HYBRIDAPPROX was much faster for instances where they both finished.

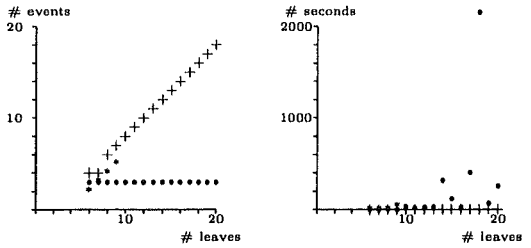


Fig. 8. Performance of algorithms for accuracy (left) and for time (right) on difficult trees to give upper bound (catepillars $(1, 2, 3, \ldots, n)$ and $(4, 5, 6, \ldots, n, 1, 2, 3)$. Number of hybrid events reported by RIATA-HGT (+), HYBRIDNUMBER (o*), and EEEP (*).

This suggests that an FPT result might be possible for the MINIMUM HYBRIDIZATION problem.

The gap between the upper and lower bound results leave much room for improvement in estimating the hybridization number. Based on the analysis in [7, 19], it appears that the upper bound heuristics work well in practice. However, there are instances (e.g. the simulated caterpillar dataset) for which they significantly overestimates the exact answer. Inferring from the caterpillar dataset, this could be that the upper bounds are high, but more work is needed to determine that. Despite the simplicity of the AVERAGECHERRIES heuristic, it gives linear-time lower bounds that do well in our simulated datasets. This suggests that heuristics based on more sophisticated dissimilarity in structure could yield even better results.

7 Conclusion & Future Work

We give new lower bounds for the hybridization number that do well and complement the previous work on upper bounds. Using recent theoretical work, we developed a proof-of-concept software for exact algorithms for calculating the hybridization number and rSPR distance between rooted, binary phylogenetic trees. This software showed the effectiveness of the algorithms and points to routines that can be improved in running-time and memory management to expand the size of problems that can be solved. Future work includes developing better heuristics and improving the exact algorithms to work on the increasingly large instances provided by biological data.

Acknowledgments

We would like to thank Rob Beiko, Nick Hamilton, and Daniel Huson for helpful conversations, as well as Luay Nakhleh for the RIATA-HGT code. We thank Heiko Schmidt for providing us with the reconstructed gene trees of the grass dataset and the following for use of their computational clusters: the Allan Wilson Centre, the Department of Bioinformatics at the University of Düsseldorf, and the Research Computing Group at the City University of New York. The first author was financially supported by the Allan Wilson Centre and the Wiener Wissenschafts-, Forschungs- und Technologiefonds (WWTF). The second author was supported by the US National Science Foundation (0215942 and 0513660). The last author was supported by the New Zealand Marsden Fund (UOC310).

References

1. Addario-Berry, L., Hallett, M., and Lagergren, J. (2003). Towards identifying lateral gene transfer events. In: *Proceedings of the Pacific Symposium on Biocomputing*, Vol. 8, pp. 279-290.
2. Allen, B. L. and Steel, M. (2001). Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, **5**, 1-13.
3. Baroni, M. (2004). Hybrid phylogenies: a graph-based approach to represent reticulate evolution. Unpublished PhD thesis, University of Canterbury.
4. Baroni, M., Semple, C., and Steel, M. (2004). A framework for representing reticulate evolution. *Annals of Combinatorics*, **8**, 391-408.
5. Baroni, M., Semple, C., and Steel, M. (2006). Hybrids in real time. *Systematic Biology*, **55**, 46-56.
6. Baroni, M., Grünewald, S., Moulton, V., and Semple, C. (2005). Bounding the number of hybridization events for a consistent evolutionary history. *Mathematical Biology*, **51**, 171-182.
7. Beiko, R. and Hamilton, N. (2006). Phylogenetic identification of lateral genetic transfer events. *BMC Evolutionary Biology*, **6**:15.
8. Bonet, M. K., St. John, K., Mahindru, R., and Amenta, N. (2006). Approximating subtree distances between phylogenies. Technical Report #669, Centre de Recerca Matematica, Barcelona.
9. Bordewich, M. and Semple, C. (2004). On the computational complexity of the rooted subtree prune and regraft distance. *Annals of Combinatorics*, **8**, 409-423.
10. Bordewich, M. and Semple, C. Computing the minimum number of hybridisation events for a consistent evolutionary history, submitted.

11. Ellstrand, N. C., Whitkus, R., and Rieseberg, L. H. (1996). Distribution of spontaneous plant hybrids. *Proc. Natl. Acad. Sci.*, **93**:10, 5090–3.
12. Grass Phylogeny Working Group (2001). Phylogeny and subfamilial classification of the grasses (*poaceae*). *Ann. Mo. Bot. Gard.*, **88**:3, 373–457.
13. Hallett, M. and Lagergren, J. (2001). Efficient algorithms for lateral gene transfer problems. In: *Proceedings of the Fifth Annual International Conference on Research in Computational Molecular Biology* (RECOMB 2001), ACM Press, New York, pp. 149-156.
14. Hein, J., Jing, T., Wang, L., and Zhang, K. (1996). On the complexity of comparing evolutionary trees. *Discrete Applied Mathematics*, **71**, 153-169.
15. Maddison, W. (1997). Gene trees in species trees. *Systematic Biology*, **46**, 523-536.
16. McGeoch, C.C. (1992). Analyzing algorithms by simulation: variance reduction techniques and simulation speedups. *ACM Comp. Surveys* **24**, 195–212.
17. Moret, B.M.E. (2002). Towards a discipline of experimental algorithmics. In: *Proc. 5th DIMACS Challenge* (ed: M.H. Goldwasser, D.S. Johnson, and C.C. McGeoch), DIMACS Monographs **59**, 197-213, American Mathematical Society, Providence, 2002.
18. Munzner, T., Guimbrètiere, F., Tasiran, S., Zhang, L., and Zhou, Y. (2003). TreeJuxtaposer: Scalable tree comparison using Focus+Context with guaranteed visibility. *SIGGRAPH 2003 Proceedings, published as special issue of Transactions on Graphics*, 453–462, 2003.
19. Nakhleh, L., Ruths, D., and Wang, L. S. (2005). RIATA-HGT: a fast and accurate heuristic for reconstructing horizontal gene transfer. In: *Proceedings of the Eleventh International Computing and Combinatorics Conference* (COCOON 05) (ed. L. Wang), Lecture Notes in Computer Science, Vol. 3595, Springer, pp. 84-93.
20. Nakhleh, L., Warnow, T., Linder, C. R., and St. John, K. (2005). Reconstructing reticulate evolution in species—theory and practice. *Journal of Computational Biology*, **12**, 796-811.
21. Olsen, G. J., Matsuda, H., Hagstrom, R., and Overbeek, R. (1994). fastDNAmL: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci.*, **10**:1, 41–8.
22. Rieseberg, L. H., Raymond, O., Rosenthal, D. M., Lai, Z., Livingstone, K., Nakazato, T., Durphy, J. L., Schwarzbach, A. E., Donovan, L. A., and Lexer, C. (2003). Major ecological transitions in wild sunflowers facilitated by hybridization. *Science*, **301**:5637, 1211–6.
23. Rodrigues, E. M., Sagot, M. -F., and Wakabayashi, Y. (2001). Some approximation results for the maximum agreement forest problem. In: *Approximation, Randomization and Combinatorial Optimization: Algorithms and Techniques* (APPROX and RANDOM) (ed. M. Goemans *et al.*), Lecture Notes in Computer Science, Vol. 2129, Springer, Berlin, pp. 159-169.
24. Sanderson, M. J. (2003). r8s; inferring absolute rates of evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, **19**:301-302.
25. Schmidt, H. A. (2003). Phylogenetic trees from large datasets. PhD thesis, Heinrich-Heine-Universität, Düsseldorf, Germany.
26. Semple, C. (2006). Hybridization networks. In: *New Mathematical Models of Evolution* (ed. O. Gascuel and M. Steel), Oxford University Press, in press.
27. Semple, C. and Steel, M. (2003). *Phylogenetics*. Oxford University Press.